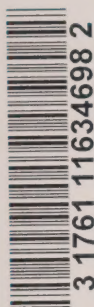


CA1  
BS1  
-1990  
R32

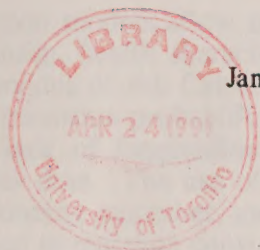


SMOOTHING PROCEDURES FOR SIMULATED  
LONGITUDINAL MICRODATA

by

Jane F. Gentleman, Dale Robertson  
and Monica Tomiak

No. 32



Statistics Canada  
Analytical Studies Branch

# Research Paper Series



Statistics  
Canada

Statistique  
Canada

Canada



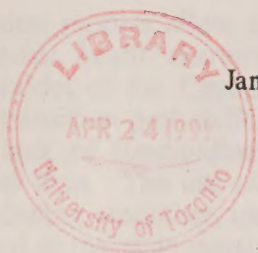
CAI  
BSI  
-1990  
R32

**SMOOTHING PROCEDURES FOR SIMULATED  
LONGITUDINAL MICRODATA**

by

Jane F. Gentleman, Dale Robertson  
and Monica Tomiak

No. 32




Social and Economic Studies Division  
Analytical Studies Branch  
Statistics Canada  
1990

The analysis presented in this paper is the responsibility of the authors and does not necessarily represent the views or policies of Statistics Canada.

Aussi disponible en français





Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto

<https://archive.org/details/31761116346982>

# Smoothing Procedures for Simulated Longitudinal Microdata

Jane F. Gentleman, Dale Robertson, and Monica Tomiak

Statistics Canada

Key Words: Longitudinal Data, Microsimulation Model, Simulation, Smoothing

## ABSTRACT

Microsimulation models allow one to study the behavior of a large population over time. At Statistics Canada, health characteristics and risk factors are being added to a demographic and labor force model of the Canadian population. This paper describes a method for obtaining multivariate transition probabilities between states for use in advancing individuals in simulated time. The lack of longitudinal data means that these probabilities must be derived from cross-sectional data. The use of transition probabilities by the microsimulation model has the effect of producing smoother, more realistic, logically possible life histories. The probabilities are constrained to maintain consistency with the cross-sectional distributions. The constraints on the probabilities may be expressed as those of the transportation problem in network flow theory. The objective function in this special type of linear program is chosen to discourage unrealistically large or frequent changes of state across time. Canada Health Survey data were used to generate multivariate transition probability arrays for smoking, blood pressure, cholesterol, and body mass index, all thought to be important risk factors for coronary heart disease.

## 1. INTRODUCTION

This paper describes techniques for enabling a dynamic microsimulation model which relies on cross-sectional source data to nevertheless produce realistically smooth simulated longitudinal microdata. A microsimulation model consists of a set of algorithms and a computer program which simulate microdata. The algorithms are based on probabilistic or deterministic submodels, and/or on observed distributions of real data. The microsimulation model generates a sample of simulated units which represent some conceptual population of units. These units might, for example, be people, households, or business firms. We shall refer to them as "individuals". The sample of individuals is used to make inferences about the population.

Received: February 16, 1990

Accepted: August 21, 1990



Microsimulation models are particularly useful for posing and answering questions of a "what if" nature. To distinguish the data used in the construction of a microsimulation model from the data generated by such a model, we shall refer to the former as "source data" and to the latter as the "simulated data" or "sample data".

Since the 1960's, microsimulation models have been used for the analysis of public policy (for evaluating social security or other government social programs, income tax, etc.) and in the fields of demography, economics, energy, health, etc. For a useful, broad collection of papers concerning microsimulation, see Orcutt, Merz, and Quinke (1986).

A dynamic microsimulation model ages a sample of individuals across time, simulating multivariate data (such as marital status, employment status, education, consumption of manufactured goods, and health status) which describe them during each time period. There exist many panel and other surveys which can provide source data which are both multivariate and longitudinal, but the need of a microsimulation model for such data often cannot be fully met. Hoschka (1986, p. 49) lists "missing variables" and "cross section instead of panel surveys" as being among the most common shortcomings of microsimulation model source data. By their very nature, longitudinal data require a long period of time to be collected, and it is not always possible to foresee what combinations of variables will be needed, so that alternate strategies are needed. Indeed, the ability to produce unforeseen combinations of (admittedly synthetic) variables is one of the strengths of microsimulation.

Assume that for each variable of interest, a finite number of outcomes (or classes, or states) have been defined. From longitudinal age-specific source microdata, it is possible to

estimate the distribution of a variable at a given age  $t$ , and to estimate transition probabilities for an individual moving from a certain state at age  $t$  to a certain state at age  $t+1$ . These probabilities can be used by the microsimulation model as it ages the sample.

In the absence of longitudinal source data, analysts often use cross-sectional data, treating the age-specific source data gathered at one point in time as if it were data describing one group of individuals across time. Elandt-Johnson (1980) examines the relationship between cross-sectional and longitudinal data in survival analysis, concluding that "if there is a relationship between a characteristic and age for an individual, it cannot be uniquely estimated from population cross-sectional data, unless its functional contribution to the hazard rate is determined...Longitudinal studies are necessary to obtain more reliable information."

Transition probabilities cannot in general be deduced from cross-sectional data (a deficiency which also occurs with longitudinal data which are collected, but not linked, across time). However, if a microsimulation model ignores transitions and generates data independently for each age, the characteristics of a simulated individual may vary unrealistically across time, even though the distribution of the sample matches the source data distribution at each age.

For example, suppose that cross-sectional source data were used to estimate at each age the distribution of a variable describing an individual's smoking habit (classed as "Never Smoker", "Current Smoker", or "Former Smoker"). If the microsimulation model generates an individual's smoking habit independently for each age, the resulting simulated smoking history may have unrealistically frequent changes of state, and it may exhibit a logically impossible



transition (such as from being a current smoker to being a never smoker). Ideally, a microsimulation model would use an array of multivariate transition probabilities to move an individual from one age to the next.

In the absence of multivariate source data, analysts may "synthetically" link different data files (enhancing the data for one individual by appending data from another, similar, individual), and they may resort to assuming independence of separate variables. In the latter case, multivariate transition probabilities are simply products of univariate transition probabilities, so they are easily calculated and require relatively little computer storage space. On the other hand, if the variables are not independent and their joint distribution is available, then the number of combinations of variables and states can become prohibitively large. Orcutt (1986, p. 19) describes the storage problem:

"Even an extremely modest microanalytic model of an economy involving persons embedded within families would result in substantially more than ten endogenous variables per family...And, even if only ten values were permitted for each variable, the number of cells needed to classify families without loss of information would be ten billion. The full matrix of transition probabilities would then have ten billion squared elements!"

It may therefore be necessary to assume that small groups of variables are independent of other small groups of variables.

Krupp (1986, p. 36) discusses the computational demands made by microsimulation models:

"If one assumes that a simulation is based on 20,000 households, that updating a characteristic of a household or producing a behavioral change requires an average of about 20 operations at the programming language level, and that for every household approximately 100 socially relevant characteristics are considered, then the simulation for one year demands 20 million operations at the programming language level. A simulation over ten years requires 200 million operations and over fifty years one billion."



Other problems that arise as the number of variables and/or states increases are the usual statistical ones of small sample sizes, and the loss of observations due to missing data ("item nonresponse").

This paper describes procedures for obtaining multivariate transition probability arrays from cross-sectional multivariate data (or from unlinked longitudinal data) in order to smoothe the longitudinal behavior of the simulated individuals. The examples provided utilize data for four variables having 5, 3, 3, and 4 states, respectively, so that there are 180 frequencies at each age group, and  $180 \times 180 = 32,400$  transition probabilities from one age group to the next. Across the 12 age groups, there are therefore 356,400 transition probabilities. Given multivariate data for two adjacent age groups, an array of multivariate transition frequencies (and the corresponding array of transition probabilities) is obtained using linear programming (LP) methods. These transition frequencies are made consistent with the cross-sectional multivariate source data, and conditions which are innate to the particular variables are also imposed; these, plus the nonnegativity of the frequencies, form the constraints of the linear program. The linear program's objective function is chosen so that transitions to "nearby" states are favored over transitions to "distant" states (which is reasonable if the time interval between the two age groups is relatively small).

The approach here is from a smoothing rather than an estimation point of view because of the very large number of degrees of freedom available for determining transition frequencies given relatively few marginal sums. Our approach is analogous to that used in smoothing ordinary univariate time series data, for which there are many possible smoothing algorithms (see, e.g., Dagum (1985), Cleveland and Kleiner (1975), Velleman and Hoaglin (1981, Chapter 6), and Cleveland (1985, pp. 167-178)); the choice of an algorithm and the parameter values

thereof is often made heuristically, in order to obtain the desired quality and degree of smoothing. It is in that spirit that procedures are proposed here for generating realistically smooth longitudinal microdata.

The use of LP methods to generate  $4 \times 4$  transition frequency matrices - with the reverse objective of maximizing mobility by favoring transitions to distant states - is described in Meyer (1978). The estimation of a  $2 \times 2$  transition matrix in the context of generalized linear models is discussed in McCullagh and Nelder (1983, pp. 175-177).

Problems of estimation using observed transition frequencies (called "gross flows" in the context of labor force panel data) were addressed in a 1984 conference devoted to that topic (see U.S. Dept. of Commerce and U.S. Dept. of Labor (1985)).

This study was motivated by the need of the POHEM health microsimulation model (see Wolfson (1989)), which is being developed at Statistics Canada, to produce plausibly smooth simulated data. An existing model (see Wolfson (1989a, pp. 30-34)) already generates and ages over time a sample of individuals to whom are assigned demographic and labour force characteristics typical of Canada. Individual health histories, health risk factor exposures, and medical costs are now being added to the model. The ultimate goals in building this model are to be able to evaluate and compare different health policies, develop indices of the state of health of the Canadian population, and identify needs for better and more health data.

Section 2 below describes the techniques used to obtain transition probabilities, examines the statistical deficiencies of cross-sectional data, and considers the effect of unobserved heterogeneity in mortality rates. Section 3 gives examples using real data from the Canada



Health Survey. Section 4 considers computational efficiency and discusses the possibility of rephrasing the linear programming problem in two alternative mathematical forms (as a network flow problem and as a transportation problem).

## 2. THE SMOOTHING TECHNIQUE

### 2.1 Definition of Terms

Suppose that there are  $k$  variables of interest ( $k \geq 1$ ), for which multinomial data are available as follows: For each variable, a finite set of mutually exclusive, exhaustive possible outcomes (states) has been defined, and cross-classified frequencies of occurrence of each outcome combination have been observed for  $n_t$  individuals of age  $t$  and for  $n_{t+1}$  individuals of age  $t+1$ . If the data are cross-sectional, these two groups of individuals are disjoint, and  $n_{t+1}$  may even be larger than  $n_t$  (which cannot occur in a closed population). It will be assumed here that the observed proportions of individuals in each state at age  $t$  and at age  $t+1$  are representative of those which would have been observed for one cohort of individuals at two adjacent ages.

The array of transition frequencies between age  $t$  and age  $t+1$  (and the corresponding array of transition probabilities) is  $2k$ -dimensional. For notational simplicity,  $k$  will be assumed to be equal to 2, without loss of generality. Suppose, then, that the number of states for Var. 1 is  $s_1$ , and the number of states for Var. 2 is  $s_2$ . Let  $u_{i_1 i_2}$  be the number of individuals who were observed to be in the bivariate state  $(i_1, i_2)$  at age  $t$ , and let  $v_{j_1 j_2}$  be the observed number of individuals in state  $(j_1, j_2)$  at age  $t+1$ . (Here  $i_1$  and  $j_1$  label the state for Var. 1, and  $i_2$  and  $j_2$  label the state for Var. 2;  $i_1$  and  $j_1 = 1, \dots, s_1$ ;  $i_2$  and  $j_2 = 1, \dots, s_2$ ). Then  $n_t = u_{..}$  and  $n_{t+1} = v_{..}$  (where

the dot notation signifies summation over the indicated subscript). Ordinarily,  $n_t \neq n_{t+1}$ ; this occurs in a closed population because of losses due to mortality, and in cross-sectional data because the two groups contain different individuals.

For the time being, assume that there is no mortality between the two ages, and rescale the observed frequencies (for either or both ages) so that the number ( $n$ ) of individuals represented at each age is the same: Multiply the  $u_{i_1 i_2}$ 's by a constant  $C$  and the  $v_{j_1 j_2}$ 's by  $C \frac{n_t}{n_{t+1}}$ . For example, multiply each observed frequency at age  $t$  by  $C = \frac{n_{t+1}}{n_t}$ , in which case the frequencies at age  $t+1$  remain unchanged. Since the two sets of rescaled frequencies now have a common sum, the quantities  $\left\{ \frac{n_{t+1}}{n_t} u_{i_1 i_2} \right\}$  and  $\{v_{j_1 j_2}\}$  can be treated as the marginal sums  $\{x_{i_1 i_2 \dots}\}$  and  $\{x_{\dots j_1 j_2}\}$ , respectively, of the array  $\{x_{i_1 i_2 j_1 j_2}\}$  of unknown transition frequencies which are to be determined using LP methods. As discussed below, the resulting transition probabilities are invariant to the choice of the scale factor  $C$ . The assumptions implicit in performing such rescaling are discussed in Section 2.2.

The transition frequency  $x_{i_1 i_2 j_1 j_2}$  is the unknown number of individuals who made the transition from state  $(i_1, i_2)$  at age  $t$  to state  $(j_1, j_2)$  at age  $t+1$ . The overall sum of the transition frequencies is  $x_{\dots} = C n_t = n$ . The transition probability  $p_{i_1 i_2 j_1 j_2}$  is the probability of an individual being in state  $(j_1, j_2)$  at age  $t+1$ , conditional on having been in state  $(i_1, i_2)$  at age  $t$ :

$$p_{i_1 i_2 j_1 j_2} = \frac{x_{i_1 i_2 j_1 j_2}}{x_{i_1 i_2 \dots}}. \quad (1).$$

(The word "probability" is used informally throughout the discussion here, and may instead be interpreted as "proportion", in which case, the mortality rates discussed in section 2.2 are estimates of mortality rates.) The goal is to obtain reasonable values for the  $p_{i_1 i_2 j_1 j_2}$ 's (or



equivalently for the  $x_{i_1 i_2 j_1 j_2}$ 's) for use in generating microsimulated data.

Using standard linear programming techniques, values  $\{x_{i_1 i_2 j_1 j_2}\}$  are determined so as to minimize an objective function, which is a weighted sum of the  $x_{i_1 i_2 j_1 j_2}$ 's, subject to the following three types of constraints: (i) The frequencies must be non-negative:

$$x_{i_1 i_2 j_1 j_2} \geq 0 \quad \forall i_1, i_2, j_1, j_2 \quad (2);$$

(ii) The marginal sums of the input multinomial data must be maintained :

$$\sum_{i_1} \sum_{i_2} x_{i_1 i_2 j_1 j_2} = x_{\cdot \cdot j_1 j_2} \quad \forall j_1, j_2$$

and (3);

$$\sum_{j_1} \sum_{j_2} x_{i_1 i_2 j_1 j_2} = x_{i_1 i_2 \cdot \cdot} \quad \forall i_1, i_2$$

and (iii) Relationships innate to the variables must be maintained (e.g., that the number of transitions from being a current smoker to being a never smoker is zero).

The weights for the objective function are chosen here to favor stability by discouraging transitions to distant states, assuming that the concept of "distance" between states is meaningful. With the state labels suitably ordered, a reasonable choice for the weight  $w_{i_1 i_2 j_1 j_2}$  for  $x_{i_1 i_2 j_1 j_2}$  might be a measure of the distance between the state  $(i_1, i_2)$  at age  $t$  and the state  $(j_1, j_2)$  at age  $t+1$ , such as  $|i_1 - j_1| + |i_2 - j_2|$  or  $(i_1 - j_1)^2 + (i_2 - j_2)^2$ .

There remains the question of which variables to use: the transition frequencies or the transition probabilities. That is, the result of minimizing

$$z = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w_{i_1 i_2 j_1 j_2} x_{i_1 i_2 j_1 j_2} \quad (4)$$

is in general different from the result of minimizing

$$z' = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w_{i_1 i_2 j_1 j_2} p_{i_1 i_2 j_1 j_2} = \sum_{i_1} \sum_{i_2} \sum_{j_1} \sum_{j_2} w'_{i_1 i_2 j_1 j_2} x_{i_1 i_2 j_1 j_2} \quad (5)$$

(where  $w'_{i_1 i_2 j_1 j_2} = \frac{w_{i_1 i_2 j_1 j_2}}{x_{i_1 i_2 \dots}}$ ). Using fixed weights on the frequencies tends to give more weight to the

more populous states, while using fixed weights on the probabilities tends to weight all transitions more equally, regardless of the proportion of individuals actually in those states. Comparisons of results using the two approaches are given in Section 3. In either case, the problem is formulated using frequencies as the variables. The marginal sums are then integers, and the solutions - the numbers of people making transitions - are then also integers (see Section 4).

Meyer (1978) maximized mobility by applying fixed weights to frequencies rather than to transition probabilities.

The LP method applied to the same observed data with two different choices of the rescaling factor  $C$  will yield the same array of transition probabilities (but not of transition frequencies) in both cases. If  $\{x_{i_1 i_2 j_1 j_2}\}$  are the transition frequencies resulting from a choice of  $C = C_1$ , then it is straightforward to show that the transition frequencies resulting from an alternative choice of  $C = C_2$  are  $\left\{ \frac{C_2}{C_1} x_{i_1 i_2 j_1 j_2} \right\}$ . Both transition frequency arrays have the same transition probabilities.

## 2.2 The Missing Mortality Variable

It is instructive to interpret mortality as an additional variable for which two states - Alive and Dead - are defined at any given time. (Dead is an absorbing state; persons who are dead



remain "forever" in the same multivariate state, and age for them is interpreted as the number of years since birth.) Viewed the other way around - as a life table to which variables representing other means of transition than dying have been added - the transition frequencies are similar to entries in a multistate life table (see, e.g., Rogers (1980)). In longitudinal data for a closed population, the numbers of individuals Alive and Dead are known at any given time, and transition frequencies between the two states are known; in cross-sectional data, the number Alive, but not the number Dead, is known, and no transition frequencies are known.

To illustrate this, assume, first, that the population is closed, and let Var. 1 be an ordinary variable with 3 states, and Var. 2 be Mortality (with Alive as state 1 and Dead as state 2). Then  $n_t$  is the number still alive at age  $t$  plus the number who died before age  $t$ , and  $n_t = n_{t+1} = n$ , so that no rescaling is necessary. Therefore,  $u_{i_1 i_2} = x_{i_1 i_2 \dots}$  and  $v_{j_1 j_2} = x_{\dots j_1 j_2}$ . Because Dead is an absorbing state, some transition frequencies are identically zero :  $x_{i_1 2 j_1} = 0$  (for all  $i_1, j_1$ ), and  $x_{i_1 1 j_1 2} = x_{i_1 2 j_1 2} = 0$  (for  $i_1 \neq j_1$ ). Therefore, certain other transition frequencies can be deduced from the observed marginal sums. For example  $x_{i_1 1 i_1 2}$ , which is the number of individuals who were alive and in state  $i_1$  at age  $t$  and dead (and therefore still in state  $i_1$ ) at age  $t+1$ , can be written as

$$x_{i_1 1 i_1 2} = x_{\dots i_1 2} - x_{i_1 2 \dots} \quad (6).$$

Also, the number who died during the age interval  $[t, t+1)$  is  $x_{\dots 1 \dots} - x_{\dots 2 \dots}$ , which is equal to  $x_{\dots 1 \dots} - x_{\dots 2 \dots}$ . From these quantities, the overall mortality rate  $q_t$  and the "state-specific" mortality rates  $q_{i_1}$  can be calculated :

$$q_t = \frac{(x_{\dots 1 \dots} - x_{\dots 2 \dots})}{x_{\dots 1 \dots}} = \frac{(x_{\dots 1 \dots} - x_{\dots 2 \dots})}{x_{\dots 1 \dots}} \quad (7),$$

and

$$q_{ii_1} = \frac{(x_{..i_12} - x_{i_12..})}{x_{i_11..}} \quad (8).$$

The overall mortality rate is the probability of dying in  $[t, t+1)$ , conditional on surviving to age  $t$ . The state-specific mortality rate for state  $i_1$  is the probability of dying in  $[t, t+1)$ , conditional on surviving to age  $t$  and being in state  $i_1$  at age  $t$ . The overall mortality rate is the weighted average - with the  $x_{i_11..}$ 's as weights - of the state-specific mortality rates.

Continue to assume that the population is closed, but now suppose that only those Alive at age  $t$  and those Alive at age  $t+1$  were counted (which is exactly what happens with cross-sectional data). Then the observed data consist only of  $x_{i_11..}$  (for  $i_1 = 1, 2, 3$ ) and  $x_{..j_11}$  (for  $j_1 = 1, 2, 3$ ). From this information alone, the overall mortality rate can be deduced (using Eqn. 7), but the state-specific mortality rates cannot be deduced (because the counts of dead people in the numerator of Eqn. 8 are unknown). In fact, both  $n_t$  and  $n_{t+1}$  are unknown. Suppose that the data for Alive individuals were treated as observed data for Var. 1, ignoring Var. 2 (Mortality), and rescaled to achieve a common sum using scale factor  $C = \frac{x_{..1}}{x_{1..}}$ . We then have two sets of marginal sums from which transition frequencies  $y_{i_1j_1}$  can be obtained using the LP method. These can be displayed as a matrix, with rows and columns representing the states of Var. 1 at age  $t$  and at age  $t+1$ , respectively:

$y_{11}$	$y_{12}$	$y_{13}$	$y_{1.} = \frac{x_{..1}}{x_{1..}} x_{11..}$	(9).
$y_{21}$	$y_{22}$	$y_{23}$	$y_{2.} = \frac{x_{..1}}{x_{1..}} x_{21..}$	
$y_{31}$	$y_{32}$	$y_{33}$	$y_{3.} = \frac{x_{..1}}{x_{1..}} x_{31..}$	
$y_{.1}$	$y_{.2}$	$y_{.3}$	$y_{..} = x_{..1}$	
"	"	"		
$x_{..11}$	$x_{..21}$	$x_{..31}$		



The overall sum  $x_{..1}$  is the total number of individuals alive at age  $t+1$ , and the column sums are the numbers alive in each state at age  $t+1$ . Row sum  $i_1$  is equal to  $(1 - q_i)x_{i1} \dots$ , which, if  $q_{ii_1}$  were equal to  $q_i$ , would be the number of individuals alive in state  $i_1$  at age  $t$  who remained alive at age  $t+1$  (see Eqns. 6, 7, and 8). Thus, rescaling of Alive individuals at age  $t$  is equivalent to removing from the Alive population at age  $t$  all of those individuals who are going to die by age  $t+1$ , but by applying the overall mortality rate rather than the state-specific rate. Since the LP method yields the same results regardless of the rescaling factor, rescaling implies the use of a state-independent mortality rate.

This is what occurs when cross-sectional data are rescaled; the resulting transition probabilities may be interpreted as transition probabilities from age  $t$  to age  $t+1$  for only those individuals who survived to age  $t+1$ , but under the assumption that all state-specific mortality rates are the same. A microsimulation model based on cross-sectional data can utilize these transition probabilities by applying them, just before advancing the sample from age  $t$  to age  $t+1$ , to only those individuals who have survived. Using rescaled cross-sectional data is not ideal in that one likely reason for defining different states of a variable is that the mortality rate is thought to be state-dependent. Intuitively speaking, a larger percentage of high risk individuals should die in  $[t, t+1)$ , resulting in relatively fewer of them alive at age  $t+1$ . Assuming a uniform mortality rate for all states makes it appear that some high risk individuals have moved to lower risk states.

If the mortality rates for different states are in fact heterogeneous, then a more realistic set of transition frequencies for surviving individuals would be those obtained by replacing the row sums  $(1 - q_i)x_{i1}..$  in Eqn. 9 by  $(1 - q_{ii})x_{i1}..$ , resulting, after application of the LP procedure, in the following transition frequency matrix for Var. 1:

$$\begin{array}{ccc|c}
 x_{1111} & x_{1121} & x_{1131} & x_{11 \cdot 1} = x_{11..} - x_{1112} \\
 x_{2111} & x_{2121} & x_{2131} & x_{21 \cdot 1} = x_{21..} - x_{2122} \\
 x_{3111} & x_{3121} & x_{3131} & x_{31 \cdot 1} = x_{31..} - x_{3132} \\
 \hline
 x_{\cdot 111} & x_{\cdot 121} & x_{\cdot 131} & x_{\cdot 1 \cdot 1} = x_{\cdot \cdot \cdot 1} \\
 \text{"} & \text{"} & \text{"} & \\
 x_{\cdot \cdot 11} & x_{\cdot \cdot 21} & x_{\cdot \cdot 31} & 
 \end{array} \quad (10).$$

With cross-sectional source data, the  $q_{ii}$ 's are unknown, and the preferred transition frequencies of Eqn. 10 cannot be obtained from those of Eqn. 9. A dynamic microsimulation model which relies on cross-sectional data can apply the techniques described here in the following ways: (1) Use transition probabilities derived as in Eqn. 9, recognizing that another approximation has been introduced in the model; (2) Use transition probabilities derived as in Eqn. 10, using externally-obtained estimates of state-specific mortality rates; or (3) During execution, at the end of each age interval, calculate transition probabilities derived as in Eqn. 10, using the model's own state-specific mortality rates which result from whatever algorithms the model uses to cause individuals to die.

Examples of the effects of using option 1 - assuming a uniform mortality rate in the presence of heterogeneous mortality rates - are given in Vaupel and Yashin (1985). They point



out that because of heterogeneity, selection will occur and the surviving population will differ from the original population. More research is needed, they say, on the "key question of how to tell when a population is sufficiently heterogeneous that selection matters."

Option 3 is possible, but requires a large amount of computing in addition to that already used by the microsimulation model. It is also not facilitative to the tinkering which is sometimes needed to solve the LP problem (see Section 4). Also, option 3 is only feasible in "cross-section models" (which age complete cross-sections of the sample across time), not in "longitudinal simulation models" (in which one individual is aged at a time). (See Hain and Helberger (1986) for a discussion of cross-section versus longitudinal simulation models.)

### 3. EXAMPLES

The input data used to demonstrate the smoothing technique are from the 1978/79 Canada Health Survey (CHS). This was a multistage stratified household survey of 31,668 individuals. For details of the CHS, see Statistics Canada and National Health and Welfare (1981).

For each sex, and for each of 12 age groups (15-19, 20-24, 25-29, ..., 65-69, and 70+), cross-classified frequencies were obtained for the following variables and classes:

Var. 1: Body Mass Index  $\left(\frac{kg}{m^2}\right)$

- (i) <20
- (ii) [20,25]
- (iii) (25,27]
- (iv) (27,30]
- (v) >30

Var. 2: Serum Cholesterol  $\left(\frac{mg}{dL}\right)$

- (i)  $\leq 200$
- (ii) (200,240]
- (iii) >240

Var. 3: Diastolic Blood Pressure (mmHg)

- (i) <90
- (ii) [90,105]
- (iii)  $\geq 105$

Var. 4: Smoking Habit

- (i) Never Smoker
- (ii) 1-20 cigarettes/day
- (iii) >20 cigarettes/day
- (iv) Former Smoker

Frequencies were calculated using the survey weights.

These four variables are risk factors which can be used to help predict coronary heart disease (CHD). The transition probabilities derived here are to be used in a health microsimulation model being developed by Wolfson (1989); a sub-model, constructed by Wolfson and Birkett (1989), simulates the onset and progression of coronary heart disease. The CHD sub-model was inspired by the CHD microsimulation model developed using U.S. data by Weinstein et al. (1987). The choice of body mass index as a risk factor here is based on recommendations and suggested classifications in National Health and Welfare (1988); Weinstein's model uses relative weight rather than body mass index.

Transitions involving the smoking variable have certain innate constraints. The probability of becoming a former smoker immediately after being a never smoker is zero (for a short age increment during which it is assumed that only one transition occurs). The probability of becoming a never smoker after being in any of the other three smoking categories is zero. And it may be reasonable to assume that the probability of quitting smoking is less than or equal



to the probability of resuming smoking. The appropriate elements of the transition probability array must therefore obey certain equations or inequalities. The LP approach can maintain such relationships by imposing them as additional constraints. (See further discussion in Section 4.)

Table 1 contains the cross-classified frequencies of the four variables for males between two adjacent age groups: ages 30-34 and 35-39. These data are from the Canada Health Survey; they have been weighted to represent the overall Canadian population. There were 162 and 129 survey responses for these four variables in the two age groups, respectively.

Table 2 gives a one-variable example ( $k=1$ ) of the LP procedure results. Transition frequencies and probabilities for males from age group 30-34 to age group 35-39 were calculated using the marginal distributions from Table 1 of just the Smoking Habit variable. The four smoking states were ordered as they might occur for one individual across time - from never smoker to lighter smoker (1-20 cigarettes) to heavier smoker ( $> 20$  cigarettes) to former smoker. The LP procedure was applied using different combinations of weights ( $w_{ij}=|i-j|$  or  $(i-j)^2$ ) and objective functions ( $z$  from Eqn. 4 or  $z'$  from Eqn. 5). In all four cases, the (1,4), (2,1), (3,1), and (4,1) elements of the transition matrices (involving transitions from never smoker to former smoker, and from lighter smoker, heavier smoker, and former smoker to never smoker) were constrained to be zero, but no inequality constraints were imposed for quitting smoking relative to resuming smoking.

In order for the LP procedure to obtain the results in Table 2 (and in subsequent tables), certain inconsistencies in the data of Table 1 had to be removed (due to the fact that the data are cross-sectional rather than longitudinal). These adjustments to the data are described in Section 4.

In Table 2, changing from  $z$  to  $z'$  made a difference when weights  $|i - j|$  were used, but made no difference when weights  $(i - j)^2$  were used. In general, the use of weights  $|i - j|$  permits transitions to more distant states to occur than with weights based on the squared distance (or on higher powers of the distance). In the four examples, the only transitions permitted over a distance of more than one state are from being a lighter smoker to being a former smoker (weights  $|i - j|$ , objective function  $z$ , probability .08), and from never having smoked to being a heavier smoker (weights  $|i - j|$ , objective function  $z'$ , probability .02). In the latter case, however, the one-state move from never having smoked to being a lighter smoker is less probable than the two-state move (in fact, the one-state transition is impossible), which may be unrealistic.

On the other hand, the weights  $|i - j|$  generally result in larger diagonals. The diagonals of the examples using weights  $|i - j|$  are greater than or equal to the corresponding diagonals of those using  $(i - j)^2$ . All four examples have probabilities of 1.00, which is probably unrealistically large, for the zero-state transition from former smoker to former smoker. Even so, this does not imply that a quitter of smoking will remain a quitter forever, as each age transition uses a different set of transition probabilities. In the multi-variable examples discussed below, the single variable transition matrices calculated from marginal sums of a multi-variable transition frequency array become more realistic as the number of variables increases.

The probability of resuming smoking (the sum of elements (4,2) and (4,3)) is zero in all four examples, and the probability of quitting (the sum of elements (2,4) and (3,4)) is higher - either .08 or .07. Attempts to force the probability of quitting to be lower than the probability of

resuming resulted in the LP program halting in some cases because no feasible solution exists for these data under these constraints. The problem is caused by the use of cross-sectional data and by incorrect assumptions about them, not by the LP method; see further discussion in Section 4.

It is useful to inspect the transition probabilities and to examine the effects of varying the parameters, as one does when smoothing time series. One can examine the various trade-offs among the different solutions and select the most appropriate one for the microsimulation model. The acceptability of a set of transition probabilities depends strongly on the particular set of data and on the assumptions. For example, our assumption here that transitions from never smoker to former smoker are impossible is perhaps overly stringent for five-year age intervals.

Table 3 gives a two-variable example ( $k=2$ ). Transition probabilities are provided for males from age group 30-34 to age group 35-39 for the Smoking Habit and Body Mass Index variables. Results are shown for objective function  $z$  and squared distance weights. In this example, the same additional constraints were imposed on the marginal sums of the smoking frequencies as in the examples of Table 2.

The transition probability array in Table 3 is 4-dimensional. Each of the 20 matrices in the table provides values of  $p_{i_1 i_2 j_1 j_2}$  for a fixed initial state  $(i_1, i_2)$ . The subscripts  $i_1$  and  $j_1$  index the five states of *BMI*, and the subscripts  $i_2$  and  $j_2$  index the four states of Smoking Habit. Within each matrix, one row and one column of numbers are printed in boldface and italics to highlight the probabilities for zero-state transitions for each of the two variables. That row and column intersect at the multivariate "diagonal" representing a zero-state change in both variables from age 30-34 to age 35-39. The probabilities in each matrix sum to 1.0.



The stable nature of the probabilities is evident; of the 400 transition probabilities, only 37 are non-zero, and there are only two instances ( $p_{1223} = .44$  and  $p_{5243} = .04$ ) of transitions in which both variables change. Also, there is only one instance ( $p_{3151} = .08$ ) where a variable changes by more than one state. In all of the other transitions, at most one variable changes by at most one state.

Some of the diagonal probabilities are forced by the data to be zero. For example, the CHS data contain no males of age 35-39 in the ( $BMI < 20$ , Never Smoker) category, so everyone in this state at age 30-34 must exit from it, and  $p_{1111} = 0$ , necessarily.

Similarly,  $p_{4141} = 0$ ; everyone who had been in state ( $27 < BMI \leq 30$ , Never Smoker) at age 30-34 increased his  $BMI$  and moved to the state ( $BMI > 30$ , Never Smoker) at age 35-39. Meanwhile, the state ( $27 < BMI \leq 30$ , Never Smoker) was replenished by other Never Smokers coming from the slimmer group of people in state ( $25 < BMI \leq 27$ , Never Smoker).

Table 4 provides cumulative transition probabilities from age 30-34 to age 35-39 for the full quadrivariate data set ( $k=4$ ). Of the 32,400 transition probabilities, only the 129 shown in Table 4 are non-zero. The transition probability array is sparse in part because the CHS source data were disaggregated by age and sex and because the four variables of interest were not all measured for all respondents. (For example, approximately 58% of the respondents provided blood samples from which cholesterol measurements were obtained.) Thus, there were only 162 males aged 30-34 (representing 60 of the 180 possible quadrivariate states) and 129 males aged 35-39 in the data used to produce Table 4. Sparsity is also due to the type of solutions sought by the LP algorithm. See further discussion in Section 4.

All transition frequencies out of any of the 120 empty (unobserved) initial states will be zero. (Observed marginal frequencies of zero are not altered by the rescaling procedure, and the linear programming procedure preserves observed marginal frequencies). The corresponding transition probabilities will be undefined, which is reasonable since the transition probability is conditional on an event - being in that initial state - that has not been observed.

The full quadrivariate array contains (in addition to the 129 non-zero values) 10,671 transition probabilities equal to zero, and 21,600 undefined transition probabilities.

A concise data structure similar to the representation in Table 4 can be used to store the probabilities needed for the microsimulation model. Only the defined non-zero values are kept. The probabilities are stored in cumulative form to avoid repeatedly summing probabilities each time an individual's final state is randomly selected (and to avoid problems of probabilities not adding to 1.0 due to rounding error; the last cumulative probability is set to be identically 1.0). (Although some of the consecutive cumulative probabilities in Table 4 appear to be identical, implying that the corresponding transition probabilities are zero, this is simply due to the displaying there of only two decimal places.)

The initial states  $i_1, i_2, i_3, i_4$  need not be stored, but are used to determine the location of the corresponding cumulative transition probabilities. All of the necessary information can be retained by storing only three of the columns of data shown in Table 4: (1) The "Address" column of Table 4A is a vector of length 180 (one element for each initial state) used to indirectly address the data in Table 4B. The address of the first cumulative probability corresponding to initial state  $i_1, i_2, i_3, i_4$  is in location  $i_4 + (i_3 - 1)s_4 + (i_2 - 1)s_3s_4 + (i_1 - 1)s_2s_3s_4$  of this vector (where the numbers of categories for the four variables are

$s_1 = 5$ ,  $s_2 = 3$ ,  $s_3 = 3$ , and  $s_4 = 4$ , respectively). An initial state for which transition probabilities are undefined would be assigned an address of, say, -1 (indicated in Table 4A by an asterisk). (2) The "Cumulative Probability" column of Table 4B contains the cumulative probabilities (unless they are undefined) for each initial state. In this example, this is a vector of length 129. Different initial states have, in general, different numbers of cumulative probabilities, but the search for the first cumulative probability greater than a random Uniform number will never go beyond the final cumulative probability of 1.0. (3) The "State at Age  $t + 1$ " column of Table 4B, also of length 129, contains the corresponding final states. The values  $j_1, j_2, j_3, j_4$  can be stored one per byte. (Alternatively, the index number between 1 and 180 corresponding to  $j_1, j_2, j_3, j_4$  can be stored.)

Those values indicated by an asterisk in Table 4 are assigned a missing value constant (e.g., -1.0 or the dot notation of SAS). A microsimulation model which strictly adheres to the series of transition probability arrays estimated across the various age groups will never encounter these missing values, because no individual will enter an unobserved state. However, internal adjustments and tuning within the microsimulation model may create an individual in an unobserved state, in which case the missing value would signal that there is no set of transition probabilities to move him into a new state. The microsimulation model might then generate the state at age  $t+1$  independent of the state at age  $t$ , using the multivariate distribution at age  $t+1$ . (In that case, the marginal distribution at age  $t+1$  would also need to be stored).

Table 5 shows three transition matrices for the Smoking Habit variable. These were derived, respectively, from data for one, two, and four variables, using objective function  $z$  and squared distance weights. The second and third matrices in the table are the transition



probabilities obtained from the appropriate marginal sums of 4- and 8-dimensional transition frequency arrays. Zero elements of the matrices in Table 5 are left blank to simplify comparison. (The first matrix in Table 5 also appears in Table 2B.)

In the one-variable example, an individual can either remain in the same state or move "forward" to the next state.

In the two-variable example, there is an added possibility of moving "backward" from being a former smoker to being a heavier smoker, which also results in a reduction from 1.0 to .95 of the probability of remaining a former smoker.

In the four-variable example, an additional one-state backward transition is also possible, i.e., from being a heavier smoker to being a lighter smoker. (The third one-state backward transition was forbidden.)

The possible transitions in each example are shown graphically in Figure 1.

The three transition matrices in Table 5 differ because the corresponding elements in each matrix are affected by different weights in minimizing the objective function. For example, in the one-variable case, the frequency  $x_{22}$  from which diagonal element  $p_{22}$  was computed had weight zero. In the two-variable case, the corresponding frequency  $x_{2.2}$  was the sum of both diagonal and non-diagonal elements from the four-dimensional transition frequency array, so its weight was not zero, and similarly for the four-variable case. Thus, as the number of variables increases, the probability of staying in the same Smoking Habit state remains the same or decreases. Also, the use of additional variables allows the correlation structure among them to be taken into account, and Smoking Habit is apparently not independent of the other three variables. Overall, the use of larger numbers of variables has produced more realistic lower-dimensional transition matrices.

Figure 2 shows two synthetic life histories simulated using the multivariate distributions for the four variables across all 12 age groups. In Figure 2A, the 11 8-dimensional transition probability arrays obtained using objective function  $z$  and squared distance weights were used to assign states to one individual (named "Sam" for "smooth"). In Figure 2B, only the distributions at each age were used, so that the multivariate state of this individual (named "Roy" for "rough") at age  $t$  is independent of his state at age  $t+1$ . The greater smoothness and continuity of Sam's life history is clearly noticeable. Only once does he jump from one state across an intermediate state to another state, while such leaps are quite frequent in Roy's life. Note that Roy twice experiences a forbidden transition, becoming a never smoker at age 50-54, and then becoming a former smoker at age 55-59. Roy's body mass index fluctuates unrealistically, as does his cholesterol. Sam's body mass index follows quite a believable pattern of increase to age 50-54 and then declines, similar to his cholesterol. In smoking, he shows a plausible pattern, in general gradually increasing his consumption until he quits in middle age. Roy's smoking pattern is logically impossible. For these particular individuals, either blood pressure pattern is a plausible one. Note the sensible relationships among Sam's variables over time.

#### 4. COMPUTATIONAL CONSIDERATIONS

The procedure used to obtain the transition frequencies has been described above as a linear programming one. In fact, the problem falls into a very important special class of linear programs, i.e., network flow problems, and a special case of these known as the transportation problem. The term transportation problem arose from the original interpretation as finding the least costly way to route materials from supply points to demand points (see Hitchcock (1941)). The essential constraints of a transportation problem are the imposition of known row and

column totals on the non-negative elements of a matrix. By choosing some convenient ordering for the multivariate states, one can imagine the transition frequencies to be matrix elements, with the row labels corresponding to the starting states and the column labels to the ending states. The number of individuals making a transition is clearly non-negative, and the row and column sums correspond to the total numbers of individuals in the initial and final states. We want to impose consistency with these given totals, which is precisely the transportation problem framework. The objective function in our problem imposes costs on various transitions. These are analogous to the shipping costs from one place to another in the classical application. The coefficients are such that transitions to "nearer" states are cheaper than those to more distant states.

The recognition that our linear programming problem is a network flow problem has important theoretical and practical consequences. One pleasant property is that integer valued solutions are found. If the row and column totals are integers, the algorithms will return optimal solutions with integer values, so that the number of individuals making a transition is never fractional (see Chvatal (1983) and Lawler (1976)). Finding integer valued optimal solutions is difficult for a general linear program, but it is automatic with a network flow.

Other advantages of phrasing our problem as a network flow are that network flow problems can be solved significantly faster and with less computer storage space than general linear programs. "For large scale problems, contemporary commercial linear programming codes require 50-200 times as much computer time and considerably more space for data storage than special purpose network flow algorithms." (from Bradley, Brown, and Graves (1977, p. 2)). The problems solved here with less than 400 nodes are not large ones by the standards of the field and are routinely solved by available codes. In one large application (see Barr and Turner



(1981)), a transportation problem with more than 20,000 constraints and 10,000,000 variables was solved. Microsimulation models that use the  $10^{20}$  variables described by Orcutt (see Section 1) are still unachievable and will probably be so forever, although the kinds of computations thought huge by Krupp (see Section 1) are, in fact, already quite ordinary by contemporary standards.

In the simplest transportation problem, all transitions are allowed, with no upper limit on the value of an individual frequency. In this case, providing that the row and column totals are consistent (i.e., have the same overall total), the problem always has a solution (i.e., is "feasible", in the terminology of mathematical programming). The framework does have more flexibility than this simple form implies, and this flexibility is needed for our problem. For example, if certain transitions are logically impossible, then they can be forbidden. One may also impose lower and upper bounds on various variables in the solution. This corresponds to restricting the ranges of certain transition probabilities to reflect knowledge and beliefs about what is likely. When additional constraints of this type are added, the problem may no longer be feasible. (As a simple example, if enough transitions are forbidden, it may not be possible to satisfy the demand at a particular node or nodes.) In our data, infeasibility was encountered in some cases and was always traced to the same cause: our constraints made it impossible to become a never smoker from any other state (which was reasonable), but the raw data, after rescaling to achieve consistent overall sums, had more never smokers at age  $t+1$  than at age  $t$ . The resulting infeasibility was not due to a problem with the method; it was a consequence of the use of cross-sectional data as a substitute for longitudinal data. In a sense, the rescaled data contain outliers (values inconsistent with the model) and must be adjusted to meet the logically necessary condition that the proportion of never smokers can only stay the same or decrease.

Certain types of constraints, in particular those requiring one probability to be at least as large as another probability, take one outside the pure network model. Algorithms appropriate for network problems with side constraints (see Kennington and Helgason (1980)) have been designed. We did not find it necessary to use such constraints in the present study. Our results were obtained using two SAS Procedures: LP for general linear programs (SAS Institute (1985, Chapter 5)), and NETFLOW for network flow problems (SAS Institute (1985, Chapter 6)). Another Procedure, TRANS (SAS Institute (1985, Chapter 7), specializes in transportation problems, but internal difficulties in our version, since rectified by SAS, caused us to abandon its use. TNETFLOW, a superior procedure for network flow problems (SAS Institute (1986)), has now been produced.

In general, the transition probabilities produced by these methods are sparse, in our examples because we deliberately caused sparsity by our choice of weights, and due to the methods themselves. In particular, sparseness is characteristic of a simplex-method-based linear programming solution, since roughly speaking, the algorithm works with (basic feasible) solutions which have as many zeroes as possible. For modelling on small computers, this can be an advantage, reducing the storage and computation. If the source data themselves are relatively sparse, so that many possible multivariate states have not been observed and no information is available about them, it would seem premature to spend much time "tuning" the solutions, especially before performing a sensitivity analysis of the microsimulation model in which the probabilities are to be used. If desired, however, other adjustments than changing the weights may be made to reduce sparsity, such as judiciously imposing lower bounds on the solutions, or by moving to non-basic solutions. When simple objective functions are used which have small integer coefficients, there are often multiple vertices with the same value of the objective

function, i.e., the associated graph contains zero cost cycles. Systematic identification of these cycles would allow combinations of basic solutions to be formed with more non-zero values at no cost for the objective function. That is, a weighted average of two or more solutions with the same objective function value but with zeroes in different positions may be used, since the weighted average of two transition probability arrays is still a valid transition probability array.

Figure 3A shows the observed percentages of never smokers (males only) in our Canada Health Survey data across the 12 age groups. Except for one large percentage at age 15-19 and one small one at age 50-54, the percentage across the ages is roughly linear with a negative slope, but it is not uniformly non-increasing. To adjust these data, we fitted a simple linear regression on age of the logarithm of the proportion of never smokers, omitting the above-mentioned two points. Fitted values on this line were substituted for observed data when necessary to obtain a non-increasing proportion of never smokers. The relative magnitudes of the proportions of individuals in the other smoking categories were maintained.

Figure 3B shows the observed percentages of female never smokers across the 12 age groups. These show a clearly increasing trend over time, very likely due to a cohort effect in this cross-sectional data: older women in the 1978 population were more likely to be never smokers than were younger women. This illustrates a severe conflict caused by the use of cross-sectional data in place of longitudinal data. If such data were casually subjected to the procedures described above, no solutions would be found for most age transitions. The smoothing procedure thus has a side benefit of providing a warning about certain types of inconsistencies in the input data.



## 5. CONCLUDING REMARKS

The statistical concepts of heterogeneity, selection, and multistate life tables all come together when considering the problems introduced by using cross-sectional data in place of longitudinal data. More research is needed to determine how to recognize and correct for these problems, and more longitudinal data are needed in order to sidestep them. Manton (1985, p. 18) emphasizes that long-term followup data are critically needed in order to better understand longitudinal changes and "to help us disentangle the effects of systematic mortality selection from physiological aging dynamics."

On the computing front, despite what may sometimes be insurmountably high requirements of microsimulation models for resources, there is optimism that further technological advances - such as increased processing speed, higher capacity data storage, more use of dedicated computers, and parallel processing - will allow microsimulation models to expand and improve their capabilities (see Hoschka (1986)).

## 6. ACKNOWLEDGEMENTS

The authors wish to thank Michael Wolfson for motivating this study, the results of which will be implemented in his POHEM health microsimulation model (see Wolfson (1989)).

## 7. REFERENCES

- Barr, R.S. and Turner, J.S. (1981). Microdata File Merging Through Large Scale Network Technology. *Math. Prog. Study*, **15**, 1-22.
- Bradley, G.H.; Brown, G.G.; and Graves, G.W. (1977). Design and Implementation of Large Scale Primal Transshipment Algorithms. *Management Science*, **24**, 1-34.
- Chvatal, Vasek (1983). **Linear Programming**. W.H. Freeman.
- Cleveland, William S. (1985). **The Elements of Graphing Data**. Wadsworth Advance Books and Software.
- Cleveland, W.S. and Kleiner, B. (1975). A Graphical Technique for Enhancing Scatterplots with Moving Statistics. *Technometrics*, **17**, 447-454.
- Dagum, Estela Bee (1985). Moving Averages. Article in *Encyclopedia of Statistical Sciences*, Vol. 5, 630-634. Wiley.
- Elandt-Johnson, Regina C. (1980). Some Prior and Posterior Distributions in Survival Analysis: A Critical Insight on Relationships Derived from Cross-Sectional Data. *J.R.S.S. B*, **42**, 96-106.
- Hain, Winfried and Helberger, Christof (1986). Longitudinal Simulation of Life Income. In Orcutt, et al. (1986), pp. 251-270.
- Hitchcock, F.L. (1941). The Distribution of a Product from Several Sources to Numerous Localities. *J. Mathematical Physics*, **20**, 224-230.
- Hoschka, Peter (1986). Requisite Research on Methods and Tools for Microanalytic Simulation Models. In Orcutt, et al. (1986), pp. 45-54.
- Kennington, Jeff L. and Helgason, Richard V. (1980). **Algorithms for Network Programming**. Wiley.
- Krupp, Hans-Jurgen (1986). Potential and Limitations of Microsimulation Models. In Orcutt et al. (1986), pp. 31-43.
- Lawler, Eugene L. (1976). **Combinatorial Optimization: Networks and Matroids**. Holt Rinehart and Winston.
- Manton, Kenneth G. (1985). Measurements of health and disease, a transitional perspective. *Proc. Natl. Academy of Science USA*, **82**, 3-38.
- McCullagh, P. and Nelder, J.A. (1983). **Generalized Linear Models**. Chapman and Hall.

Meyer, Garry S. (1978). On the Concept of Maximum Mobility. *Population Studies*, **32**, 355-366.

National Health and Welfare (1988). Canadian Guidelines for Healthy Weights. Cat. No. H39-134/1989E.

Orcutt, Guy (1986). Views on Microanalytic Simulation Modeling. In Orcutt, et al. (1986), pp. 9-26.

Orcutt, Guy; Merz, Joachim; and Quinke, Hermann (1986), Editors. **Microanalytic Simulation Models to Support Social and Financial Policy**. Proceedings of 1983 symposium in Bonn, Germany. North Holland.

Rogers, Andrei (editor) (1980). Essays in Multistate Mathematical Demography. Special issue of *Environment and Planning A* 12(5).

SAS Institute Inc. (1985). **SAS/OR User's Guide**, Version 5 Edition.

SAS Institute Inc. (1986). Technical Report: P-146. Changes and Enhancements to the Version 5 SAS System.

Statistics Canada and National Health and Welfare (1981). The Health of Canadians. Report of the Canada Health Survey. Catalogue 82-538E. Ottawa, Canada.

U.S. Dept. of Commerce and U.S. Dept. of Labor (1985). Proceedings of the Conference on Gross Flows in Labor Force Statistics. Washington, D.C.

Vaupel, James W. and Yashin, Anatoli I. (1985). Heterogeneity's Ruses: Some Surprising Effects of Selection on Population Dynamics. *The American Statistician*, **39**, 176-185.

Velleman, Paul F. and Hoaglin, David C. (1981). **Applications, Basics, and Computing of Exploratory Data Analysis**. Duxbury Press.

Weinstein, Milton C.; Coxson, Pamela G.; Williams, Lawrence W.; Pass, Theodore M.; Stason, William B.; and Goldman, Lee (1987). Forecasting Coronary Heart Disease Incidence, Mortality, and Cost: The Coronary Heart Disease Policy Model. *Am. J. Public Health*, **77**, 1417-1426.

Wolfson, Michael C. (1989). A System of Health Statistics: Toward a New Conceptual Framework for Integrating Health Data. Paper presented at 21st General Conference of the International Association for Research in Income and Wealth, Lahnstein, West Germany, Aug. 20-26, 1989.

Wolfson, Michael C. (1989a). Divorce, Homemaker Pensions and Lifecycle Analysis. *Population Research and Policy Review*, **8**, 25-54.



Wolfson, Michael and Birkett, Nick (1989). POHEM, Population Health Module of the System of Health Statistics (SHS): Preliminary Exploration of CHD. Paper presented at Canadian Epidemiology Research Conference, Ottawa, August 1989.

A. MALES AGED 30-34

[illegible]

TABLE 1 (CONTINUED)

## B. MALES AGED 35-39

CHOL<=200															
-----															
DIASTOLIC<90					90<=DIASTOLIC<105					DIASTOLIC>=105					
-----					-----					-----					
NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		
-----					-----					-----					
BMI<20	0	632	0	7656	0	0	0	0	0	0	0	0	0	0	
20<=BMI<=25	24597	8837	83212	35285	3001	0	0	618	0	0	0	0	0	0	
25<BMI<=27	0	16071	1882	16850	0	0	0	6957	0	0	0	0	0	0	
27<BMI<=30	3925	1404	18342	15322	2307	0	0	4108	0	0	0	0	0	0	
BMI>30	19090	16310	11190	1637	1238	0	1905	0	0	0	0	0	0	0	
-----					-----					-----					
200<CHOL<=240															
-----															
DIASTOLIC<90					90<=DIASTOLIC<105					DIASTOLIC>=105					
-----					-----					-----					
NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		
-----					-----					-----					
BMI<20	0	1229	0	0	0	0	0	0	0	0	0	0	0	0	
20<=BMI<=25	8323	18618	10883	46071	7377	0	0	0	0	0	0	0	0	0	
25<BMI<=27	847	1271	1932	26181	23408	30539	0	0	0	0	0	9603	0	0	
27<BMI<=30	0	0	10507	19765	4108	0	0	10015	0	0	0	0	0	0	
BMI>30	19598	4231	0	7253	0	0	0	0	0	0	0	0	0	0	
-----					-----					-----					
CHOL>240															
-----															
DIASTOLIC<90					90<=DIASTOLIC<105					DIASTOLIC>=105					
-----					-----					-----					
NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER		
-----					-----					-----					
BMI<20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20<=BMI<=25	15469	2154	9227	1932	0	8890	0	0	0	0	0	0	0	0	
25<BMI<=27	2040	2740	7659	26796	0	644	0	7485	0	0	0	0	4105	0	
27<BMI<=30	0	0	8720	1178	0	0	0	0	0	0	0	0	0	0	
BMI>30	3782	2226	15461	682	0	1387	0	0	0	0	0	0	0	0	
-----					-----					-----					



TABLE 2  
(A) TRANSITION FREQUENCIES FOR ONE VARIABLE CASE  
(VAR. 4 : SMOKING HABIT) FOR MALES FROM AGE GROUP 30-34 TO AGE GROUP 35-39  
ELEMENTS (1,4), (2,1), (3,1), AND (4,1) ARE CONSTRAINED TO BE ZERO

		$w_{ij} =  i - j $						$w_{ij} = (i - j)^2$				
		NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER			NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER	
Z	NEVER SMOKER	152,389	3,263	0	0	155,652		152,389	3,263	0	0	155,652
	1-20 CIG/DAY	0	111,080	24,390	11,661	147,131		0	111,080	36,051	0	147,131
	>20 CIG/DAY	0	0	161,514	0	161,514		0	0	149,853	11,661	161,514
	FORMER SMOKER	0	0	0	222,417	222,417		0	0	0	222,417	222,417
		152,389	114,343	185,904	234,078	686,714		152,389	114,343	185,904	234,078	686,714
Z'	NEVER SMOKER	152,389	0	3,263	0	155,652		152,389	3,263	0	0	155,652
	1-20 CIG/DAY	0	114,343	32,788	0	147,131		0	111,080	36,051	0	147,131
	>20 CIG/DAY	0	0	149,853	11,661	161,514		0	0	149,853	11,661	161,514
	FORMER SMOKER	0	0	0	222,417	222,417		0	0	0	222,417	222,417
		152,389	114,343	185,904	234,078	686,714		152,389	114,343	185,904	234,078	686,714

TABLE 2 (CONTINUED)  
 (B) TRANSITION PROBABILITIES FOR ONE VARIABLE CASE  
 (VAR. 4 : SMOKING HABIT) FOR MALES FROM AGE GROUP 30-34 TO AGE GROUP 35-39  
 ELEMENTS (1,4), (2,1), (3,1), AND (4,1) ARE CONSTRAINED TO BE ZERO

$w_{ij} =  i - j $				
	NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER
Z				
NEVER SMOKER	.98	.02	.00	.00
1-20 CIG/DAY	.00	.75	.17	.08
>20 CIG/DAY	.00	.00	1.00	.00
FORMER SMOKER	.00	.00	.00	.00

	NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER
Z'				
NEVER SMOKER	.98	.00	.02	.00
1-20 CIG/DAY	.00	.78	.22	.00
>20 CIG/DAY	.00	.00	.93	.07
FORMER SMOKER	.00	.00	.00	1.00

$w_{ij} = (i - j)^2$				
	NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER
NEVER SMOKER	.98	.02	.00	.00
1-20 CIG/DAY	.00	.75	.25	.00
>20 CIG/DAY	.00	.00	.93	.07
FORMER SMOKER	.00	.00	.00	1.00

	NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER
NEVER SMOKER	.98	.02	.00	.00
1-20 CIG/DAY	.00	.75	.25	.00
>20 CIG/DAY	.00	.00	.93	.07
FORMER SMOKER	.00	.00	.00	1.00

TABLE 3  
TRANSITION PROBABILITIES FOR TWO-VARIABLE CASE  
VAR. 1 : BODY MASS INDEX; VAR. 4 : SMOKING HABIT  
MALES FROM AGE GROUP 30-34 TO AGE GROUP 35-39

$$\text{USING WEIGHTS : } w_{i_1 i_2 j_1 j_2} = (i_1 - j_1)^2 + (i_2 - j_2)^2$$

$(i_1, i_2) = (\text{Var. 1, Var. 4}) \text{ AT AGE } t \text{ (30-34)}$

$(j_1, j_2) = (\text{Var. 1, Var. 4}) \text{ AT AGE } t+1 \text{ (35-39)}$

AGE t	→	NEVER SMOKER ( $i_1 = 1$ )				1-20 CIG/DAY ( $i_2 = 2$ )			
↓	AGE → t+1	NEVER SMOKER	1-20 CIG/DAY	> 20 CIG/DAY	FORMER SMOKER	NEVER SMOKER	1-20 CIG/DAY	> 20 CIG/DAY	FORMER SMOKER
	↓								
BMI < 20 ( $i_1 = 1$ )	BMI < 20	.00	.00	.00	.00	.00	.13	.00	.00
	20 <= BMI <= 25	1.00	.00	.00	.00	.00	.42	.44	.00
	25 < BMI <= 27	.00	.00	.00	.00	.00	.00	.00	.00
	27 < BMI <= 30	.00	.00	.00	.00	.00	.00	.00	.00
	BMI > 30	.00	.00	.00	.00	.00	.00	.00	.00
20 <= BMI <= 25 ( $i_1 = 2$ )	BMI < 20	.00	.00	.00	.00	.00	.00	.00	.00
	20 <= BMI <= 25	1.00	.00	.00	.00	.00	.80	.00	.00
	25 < BMI <= 27	.00	.00	.00	.00	.00	.20	.00	.00
	27 < BMI <= 30	.00	.00	.00	.00	.00	.00	.00	.00
	BMI > 30	.00	.00	.00	.00	.00	.00	.00	.00
25 < BMI <= 27 ( $i_1 = 3$ )	BMI < 20	.00	.00	.00	.00	.00	.00	.00	.00
	20 <= BMI <= 25	.20	.00	.00	.00	.00	.00	.00	.00
	25 < BMI <= 27	.48	.05	.00	.00	.00	1.00	.00	.00
	27 < BMI <= 30	.19	.00	.00	.00	.00	.00	.00	.00
	BMI > 30	.08	.00	.00	.00	.00	.00	.00	.00
27 < BMI <= 30 ( $i_1 = 4$ )	BMI < 20	.00	.00	.00	.00	.00	.00	.00	.00
	20 <= BMI <= 25	.00	.00	.00	.00	.00	.00	.00	.00
	25 < BMI <= 27	.00	.00	.00	.00	.00	1.00	.00	.00
	27 < BMI <= 30	.00	.00	.00	.00	.00	.00	.00	.00
	BMI > 30	1.00	.00	.00	.00	.00	.00	.00	.00
BMI > 30 ( $i_1 = 5$ )	BMI < 20	.00	.00	.00	.00	.00	.00	.00	.00
	20 <= BMI <= 25	.00	.00	.00	.00	.00	.00	.00	.00
	25 < BMI <= 27	.00	.00	.00	.00	.00	.00	.00	.00
	27 < BMI <= 30	.00	.00	.00	.00	.00	.02	.04	.00
	BMI > 30	1.00	.00	.00	.00	.00	.43	.51	.00



TABLE 3 (CONTINUED)

AGE t		>20 CIG/DAY ( $i_2=3$ )	FORMER SMOKER ( $i_2=4$ )
↓	AGE → t+1	NEVER SMOKER    1-20 CIG/DAY    > 20 CIG/DAY    FORMER SMOKER	NEVER SMOKER    1-20 CIG/DAY    > 20 CIG/DAY    FORMER SMOKER
	↓		
BMI< 20 ( $i_1=1$ )	BMI< 20 20<=BMI<=25 25< BMI<=27 27< BMI<=30 BMI> 30	.00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00	.00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00
20<=BMI<=25 ( $i_1=2$ )	BMI< 20 20<=BMI<=25 25< BMI<=27 27< BMI<=30 BMI> 30	.00 .00 .00 .00 .00 .00 .91 .00 .00 .00 .09 .00 .00 .00 .00 .00 .00 .00 .00 .00	.00 .00 .00 .03 .00 .00 .07 .61 .00 .00 .00 .29 .00 .00 .00 .00 .00 .00 .00 .00
25< BMI<=27 ( $i_1=3$ )	BMI< 20 20<=BMI<=25 25< BMI<=27 27< BMI<=30 BMI> 30	.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 .00 .00	.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00 .00 .00 .00
27< BMI<=30 ( $i_1=4$ )	BMI< 20 20<=BMI<=25 25< BMI<=27 27< BMI<=30 BMI> 30	.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .07 .00 .00 .00 .56 .37 .00 .00 .00 .00	.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00
BMI> 30 ( $i_1=5$ )	BMI< 20 20<=BMI<=25 25< BMI<=27 27< BMI<=30 BMI> 30	.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 1.00 .00 .00 .00 .00 .00	.00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .00 .30 .00 .00 .00 .70

TABLE 4

TRANSITION PROBABILITIES FOR FOUR-VARIABLE CASE, STORED IN CONCISE FORM  
DATA FOR MALES FROM AGE GROUP 30-34 (AGE t) TO AGE GROUP 35-39 (AGE t+1)

VAR. 1 : BODY MASS INDEX; VAR. 2 : CHOLESTEROL; VAR. 3 : BLOOD PRESSURE; VAR. 4 : SMOKING HABIT  
USING WEIGHTS :

$$w_{i_1, i_2, i_3, i_4, j_1, j_2, j_3, j_4} = (i_1 - j_1)^2 + (i_2 - j_2)^2 + (i_3 - j_3)^2 + (i_4 - j_4)^2$$

$(i_1, i_2, i_3, i_4) = (\text{Var. 1}, \text{Var. 2}, \text{Var. 3}, \text{Var. 4}) \text{ AT AGE } t \quad (30-34)$

$(j_1, j_2, j_3, j_4) = (\text{Var. 1}, \text{Var. 2}, \text{Var. 3}, \text{Var. 4}) \text{ AT AGE } t+1 \quad (35-39)$

A. Address of First Cumulative Transition Probability for Each Initial State

State at Age t	Address	State at Age t	Address	State at Age t	Address
1111	*	2131	*	3221	*
1112	*	2132	*	3222	*
1113	1	2133	*	3223	*
1114	2	2134	*	3224	*
1121	3	2211	27	3231	*
1122	5	2212	31	3232	*
1123	*	2213	32	3233	*
1124	*	2214	35	3234	*
1131	*	2221	*	3311	66
1132	*	2222	*	3312	70
1133	*	2223	36	3313	*
1134	*	2224	*	3314	*
1211	*	2231	*	3321	*
1212	7	2232	*	3322	*
1213	*	2233	*	3323	*
1214	*	2234	*	3324	73
1221	*	2311	37	3331	*
1222	*	2312	38	3332	*
1223	*	2313	41	3333	*
1224	*	2314	43	3334	*
1231	*	2321	*	4111	75
1232	*	2322	*	4112	76
1233	*	2323	46	4113	78
1234	*	2324	*	4114	81
1311	*	2331	*	4121	*
1312	*	2332	*	4122	82
1313	*	2333	*	4123	*
1314	*	2334	*	4124	*
1321	*	3111	47	4131	*
1322	*	3112	*	4132	*
1323	*	3113	*	4133	*
1324	*	3114	53	4134	*
1331	*	3121	54	4211	83
1332	*	3122	*	4212	84
1333	*	3123	*	4213	85
1334	*	3124	55	4214	88
2111	9	3131	*	4221	89
2112	10	3132	*	4222	*
2113	14	3133	*	4223	90
2114	15	3134	*	4224	92
2121	20	3211	56	4231	*
2122	23	3212	59	4232	*
2123	*	3213	61	4233	*
2124	24	3214	65	4234	93

\* UNDEFINED TRANSITION PROBABILITIES (UNOBSERVED INITIAL STATES)

TABLE 4 (CONTINUED)

State at Age t	Address
4311	*
4312	95
4313	96
4314	97
4321	*
4322	101
4323	103
4324	*
4331	*
4332	*
4333	*
4334	*
5111	*
5112	109
5113	*
5114	110
5121	*
5122	*
5123	*
5124	*
5131	*
5132	*
5133	*
5134	*
5211	112
5212	114
5213	121
5214	123
5221	124
5222	*
5223	127
5224	*
5231	*
5232	*
5233	*
5234	*
5311	*
5312	128
5313	*
5314	*
5321	*
5322	*
5323	*
5324	*
5331	*
5332	*
5333	*
5334	*

\* UNDEFINED TRANSITION PROBABILITIES (UNOBSERVED INITIAL STATES)



TABLE 4 (CONTINUED)

## B. Cumulative Transition Probabilities\* and Final States

Address	Cumulative Probability	State at Age t+1
1	1.00	1114
2	1.00	1114
3	.89	2111
4	1.00	2121
5	.17	1112
6	1.00	2113
7	.12	1212
8	1.00	2212
9	1.00	2111
10	.41	2112
11	.46	2113
12	.46	2212
13	1.00	3112
14	1.00	2113
15	.05	1114
16	.33	2113
17	.86	2114
18	.89	2214
19	1.00	3114
20	.53	2121
21	.97	2221
22	1.00	3221
23	1.00	3222
24	.04	2124
25	.58	2214
26	1.00	3124
27	.07	2111
28	.57	2211
29	.86	2221
30	1.00	2311
31	1.00	2212
32	.30	2113
33	.74	2213
34	1.00	2214
35	1.00	2214
36	1.00	3233
37	1.00	2311
38	.08	2212
39	.33	2312
40	1.00	2322
41	.80	2313
42	1.00	3313
43	.22	2214
44	.27	2314
45	1.00	3314
46	1.00	2322
47	.22	2111
48	.32	3112
49	.73	3221
50	.86	4111
51	.89	4121
52	1.00	5111
53	1.00	3114
54	1.00	4121
55	1.00	4124
56	.14	3211
57	.76	3221
58	1.00	4221
59	.08	3212
60	1.00	3222
61	.23	3113
62	.46	3213
63	.71	3214
64	1.00	3313

Address	Cumulative Probability	State at Age t+1
65	1.00	3214
66	.25	2311
67	.67	3221
68	.79	3311
69	1.00	5311
70	.46	3312
71	.94	3313
72	1.00	3322
73	.71	3324
74	1.00	3334
75	1.00	5111
76	.44	3112
77	1.00	4112
78	.67	4113
79	.99	4114
80	1.00	4124
81	1.00	4114
82	1.00	3222
83	1.00	5211
84	1.00	3222
85	.42	4213
86	.91	4214
87	1.00	4224
88	1.00	4214
89	1.00	4221
90	.51	3233
91	1.00	4124
92	1.00	4224
93	.64	3334
94	1.00	4224
95	1.00	4313
96	1.00	4313
97	.24	3314
98	.69	4214
99	.89	4314
100	1.00	5314
101	.97	3222
102	1.00	3322
103	.38	3233
104	.38	3313
105	.38	3324
106	.41	3334
107	.69	4224
108	1.00	4313
109	1.00	5112
110	.77	4114
111	1.00	5114
112	.87	5111
113	1.00	5211
114	.14	4213
115	.40	5112
116	.58	5113
117	.59	5123
118	.67	5212
119	.97	5313
120	1.00	5322
121	.71	5113
122	1.00	5214
123	1.00	5214
124	.01	5111
125	.08	5121
126	1.00	5211
127	1.00	5123
128	.97	5312
129	1.00	5313

\*UNDEFINED AND ZERO-VALUED TRANSITION PROBABILITIES ARE EXCLUDED

TABLE 5  
TRANSITION PROBABILITIES FOR SMOKING HABIT  
CALCULATED FROM ONE-, TWO-, AND FOUR-VARIABLE DATA  
(USING OBJECTIVE FUNCTION Z AND SQUARED DISTANCE WEIGHTS)

A. One-Variable Example

	NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER
NEVER SMOKER	.98	.02		
1-20 CIG/DAY		.75	.25	
>20 CIG/DAY			.93	.07
FORMER SMOKER				1.00

B. Two-Variable Example

	NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER
NEVER SMOKER	.98	.02		
1-20 CIG/DAY		.75	.25	
>20 CIG/DAY			.87	.13
FORMER SMOKER			.05	.95

C. Four-Variable Example

	NEVER SMOKER	1-20 CIG/DAY	>20 CIG/DAY	FORMER SMOKER
NEVER SMOKER	.98	.02		
1-20 CIG/DAY		.73	.27	
>20 CIG/DAY		.02	.80	.18
FORMER SMOKER			.08	.92

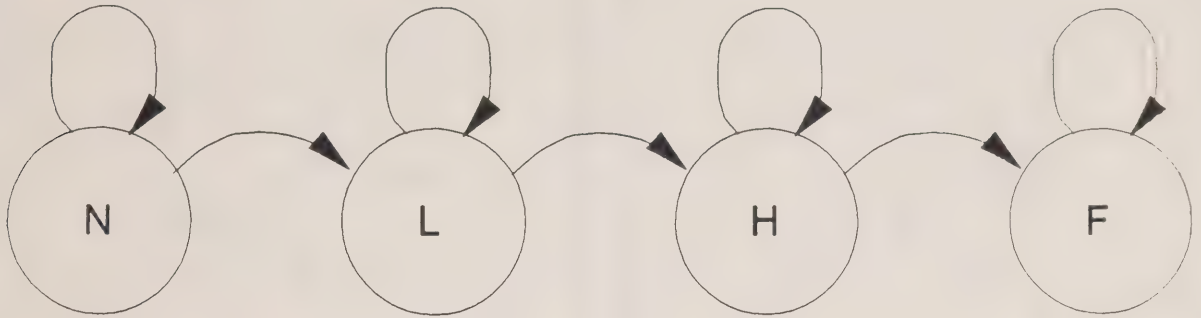
FIGURE 1

Possible Transitions for Smoking Habit

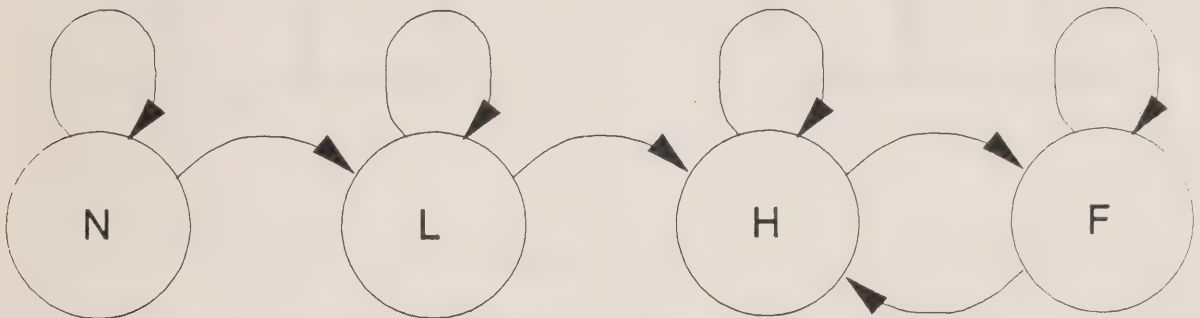
Calculated from One-, Two-, and Four-Variable Data

(using objective function Z and squared distance weights)

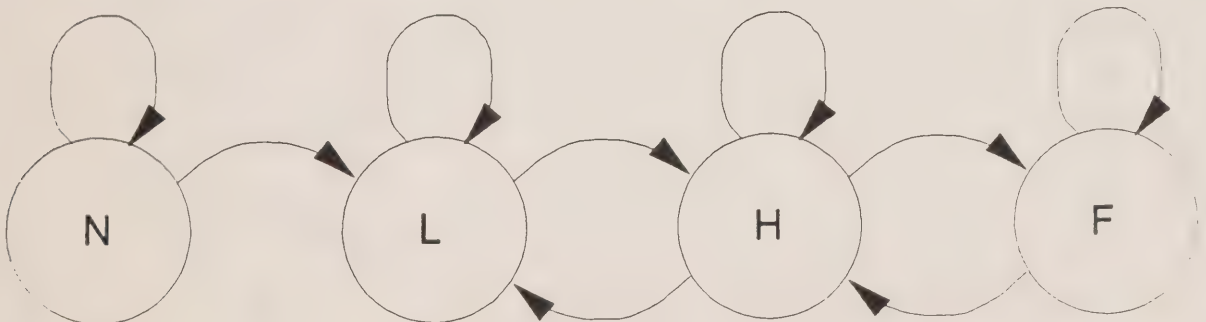
A. One-Variable Example



B. Two-Variable Example



C. Four-Variable Example



N = Never Smoker

L = Lighter Smokers (1 to 20 cigarettes per day)

H = Heavier Smokers (more than 20 cigarettes per day)

F = Former Smoker



FIGURE 2. TWO SIMULATED LIFE HISTORIES

A. SAM (Smooth history)

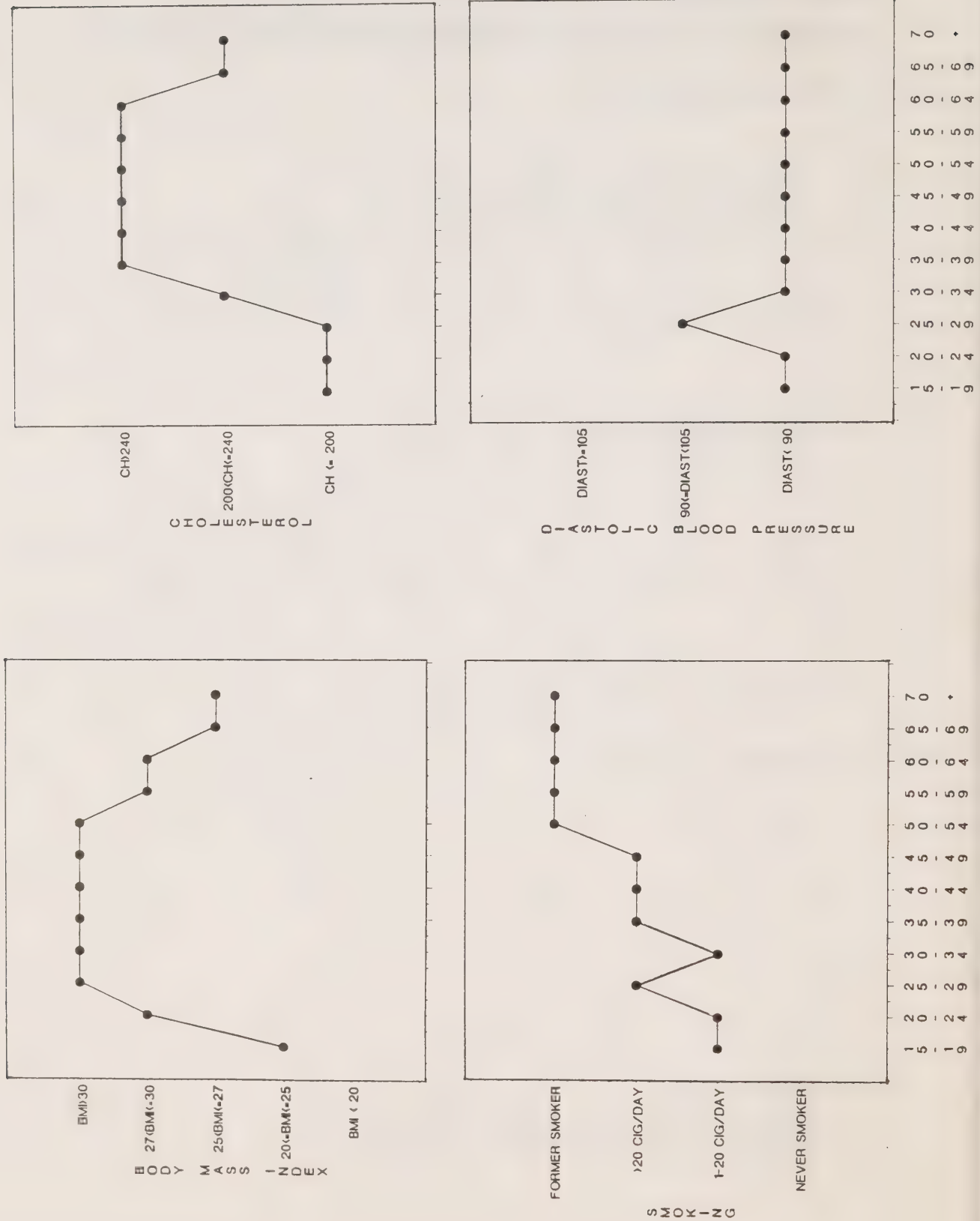


FIGURE 2 (continued).

B. ROY (Rough history)

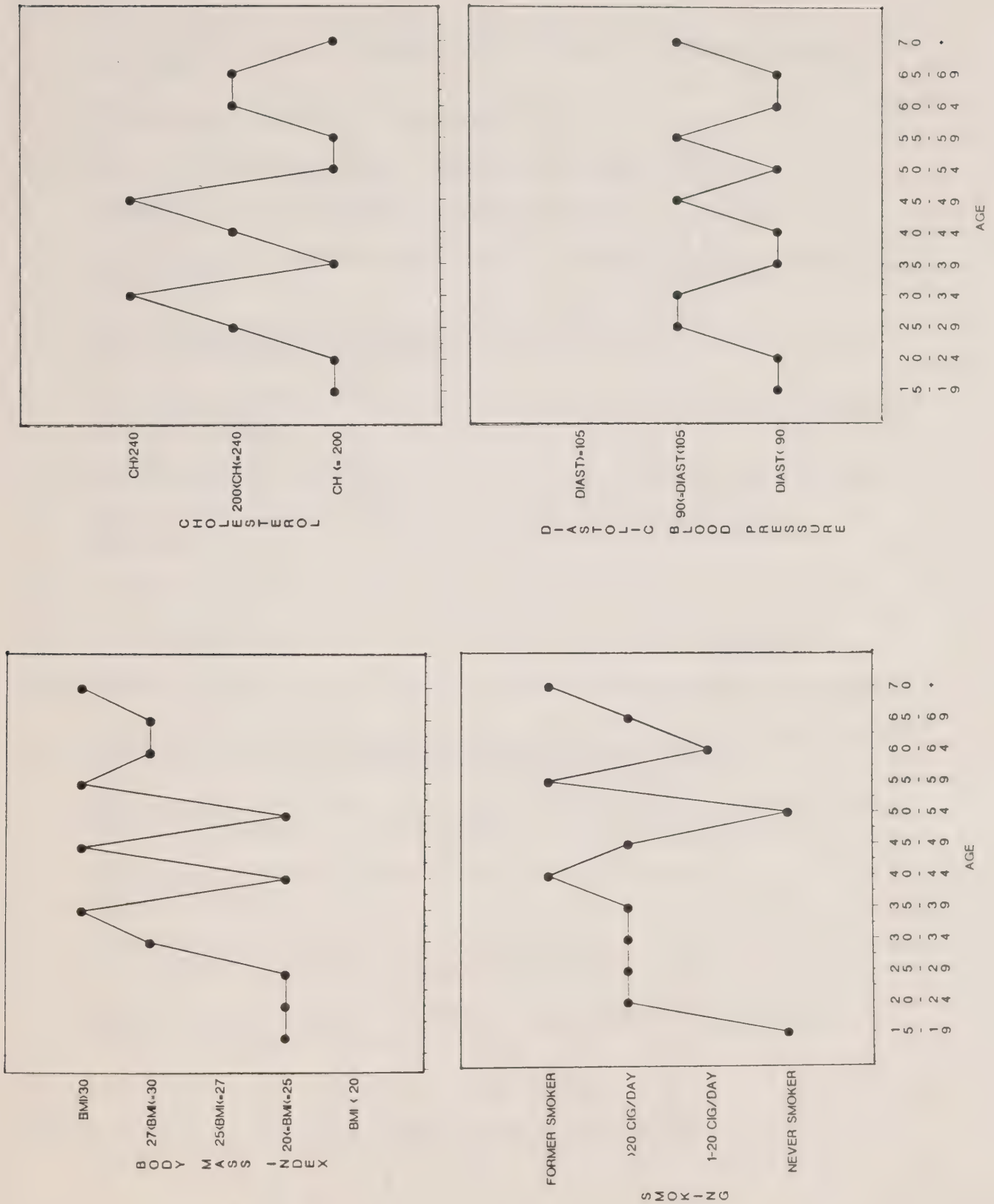
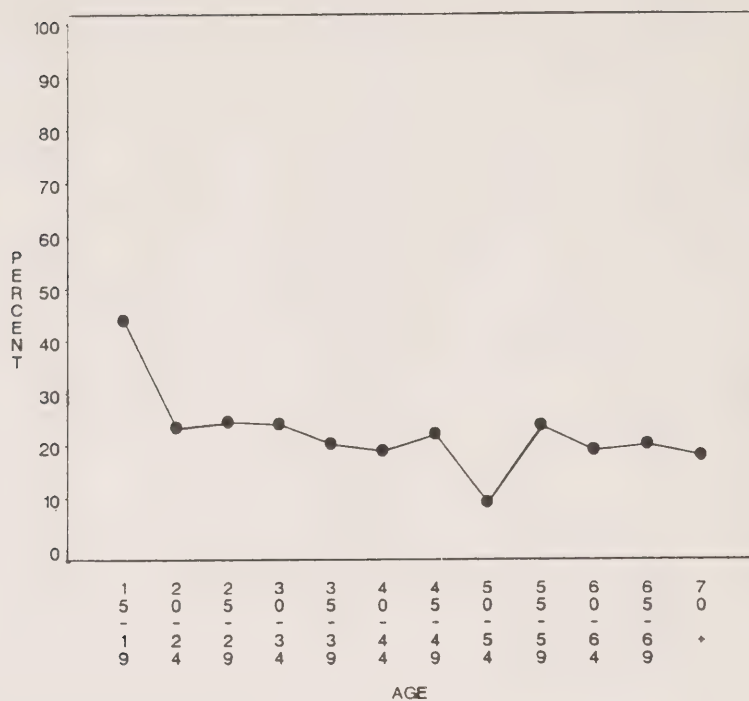
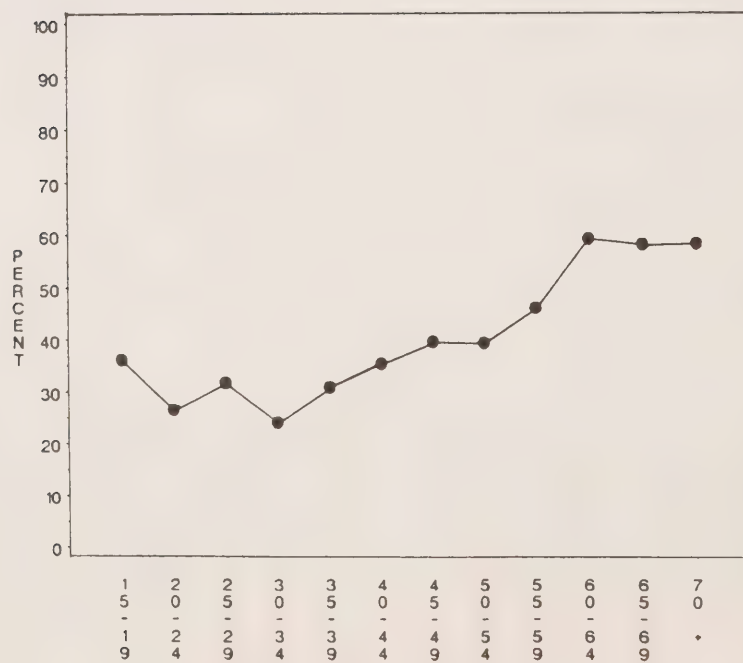


FIGURE 3. PERCENTAGES OF NEVER SMOKERS  
ACROSS 12 AGE GROUPS

A. MALES



B. FEMALES





ANALYTICAL STUDIES BRANCH  
RESEARCH PAPER SERIES

No.

1. *Behavioural Response in the Context of Socio-Economic Microanalytic Simulation*, **Lars Osberg**
2. *Unemployment and Training*, **Garnett Picot**
3. *Homemaker Pensions and Lifetime Redistribution*, **Michael Wolfson**
4. *Modelling the Lifetime Employment Patterns of Canadians*, **Garnett Picot**
5. *Job Loss and Labour Market Adjustment in the Canadian Economy*, **Garnett Picot and Ted Wannell**
6. *A System of Health Statistics: Toward a New Conceptual Framework for Integrating Health Data*, **Michael C. Wolfson**
7. *A Prototype Micro-Macro Link for the Canadian Household Sector*, **Hans J. Adler and Michael C. Wolfson**
8. *Notes on Corporate Concentration and Canada's Income Tax*, **Michael C. Wolfson**
9. *The Expanding Middle: Some Canadian Evidence on the Deskillling Debate*, **John Myles**
10. *The Rise of the Conglomerate Economy*, **Jorge Niosi**
11. *Energy Analysis of Canadian External Trade: 1971 and 1976*, **K.E. Hamilton**
12. *Net and Gross Rates of Land Concentration*, **Ray D. Bollman and Philip Ehrensaft**
13. *Cause-Deleted Life Tables for Canada (1921 to 1981): An Approach Towards Analyzing Epidemiologic Transition*, **Dhruva Nagnur and Michael Nagrodski**
14. *The Distribution of the Frequency of Occurrence of Nucleotide Subsequences, Based on Their Overlap Capability*, **Jane F. Gentleman and Ronald C. Mullin**
15. *Immigration and the Ethnolinguistic Character of Canada and Quebec*, **Réjean Lachapelle**
16. *Integration of Canadian Farm and Off-Farm Markets and the Off-Farm Work of Women, Men and Children*, **Ray D. Bollman and Pamela Smith**
17. *Wages and Jobs in the 1980s: Changing Youth Wages and the Declining Middle*, **J. Myles, G. Picot and T. Wannell**
18. *A Profile of Farmers with Computers*, **Ray D. Bollman**
19. *Mortality Risk Distributions: A Life Table Analysis*, **Geoff Rowe**



20. *Industrial Classification in the Canadian Census of Manufactures: Automated Verification Using Product Data*, John S. Crysdale
21. *Consumption, Income and Retirement*, A.L. Robb and J.B. Burbridge
22. *Job Turnover in Canada's Manufacturing Sector*, John R. Baldwin and Paul K. Gorecki
23. Series on *The Dynamics of the Competitive Process*, John R. Baldwin and Paul K. Gorecki
  - A. *Firm Entry and Exit Within the Canadian Manufacturing Sector.*
  - B. *Intra-Industry Mobility in the Canadian Manufacturing Sector.*
  - C. *Measuring Entry and Exit in Canadian Manufacturing: Methodology*
  - D. *The Contribution of the Competitive Process to Productivity Growth: The Role of Firm and Plant Turnover.*
  - E. *Mergers and the Competitive Process.*
  - F. *(in preparation)*
  - G. *Concentration Statistics as Predictors of the Intensity of Competition*
  - H. *The Relationship Between Mobility and Concentration for the Canadian Manufacturing Sector*
24. *Mainframe SAS Enhancements in Support of Exploratory Data Analysis*, Richard Johnson and Jane F. Gentleman
25. *Dimensions of Labour Market Change in Canada: Intersectoral Shifts, Job and Worker Turnover*, John R. Baldwin and Paul K. Gorecki
26. *The Persistent Gap: Exploring the Earnings Differential Between Recent Male and Female Postsecondary Graduates*, Ted Wannell
27. *Estimating Agricultural Soil Erosion Losses From Census of Agriculture Crop Coverage Data*, Douglas F. Trant
28. *Good Jobs/Bad Jobs and the Declining Middle: 1967-1986*, Garnett Picot, John Myles, Ted Wannell
29. *Longitudinal Career Data for Selected Cohorts of Men and Women in the Public Service, 1978-1987*, Garnett Picot and Ted Wannell
30. *Earnings and Death - Effects Over a Quarter Century*, Michael Wolfson, Geoff Rowe, Jane F. Gentleman and Monica Tomiak
31. *Firm Response to Price Uncertainty: Tripartite Stabilization and the Western Canadian Cattle Industry*, Theodore M. Horbulyk
32. *Smoothing Procedures for Simulated Longitudinal Microdata*, Jane F. Gentleman, Dale Robertson and Monica Tomiak
33. *Patterns of Canadian Foreign Direct Investment Abroad*, Paul K. Gorecki

For further information, contact the Chairperson, Publications Review Committee, Analytical Studies Branch, R.H. Coats Bldg., 24th Floor, Statistics Canada, Tunney's Pasture, Ottawa, Ontario K1A 0T6, (613) 951-8213.





